Introduction to Deep Generative Modeling

Lecture #2

HY-673 – Computer Science Dep., University of Crete
Professors: Yannis Pantazis & Yannis Stylianou
TAs: Michail Raptakis & Michail Spanakis

What is probability?

- <u>Frequentist Answer:</u>

Let $A \subseteq \Omega$ be an event, then:

$$P(A) = \lim_{n \to \infty} \frac{n_A}{n}, \text{ where}$$

- n: number of experiments (independent trials),
- n_A: number of occurences of event A (e.g., coin or dice).

What is probability?

- Axiomatic Answer (Kolmogorov, 1933):

Let Ω be a sample space, $\mathcal{F}$ be a event space (e.g., σ-algebra of Ω) and P be a measure. If

1. For all $A \in \mathcal{F}$: $P(A) \geq 0$.
2. $P(\Omega) = 1$.
3. $A \cap B = \emptyset \to P(A \cup B) = P(A) + P(B)$. ($\sigma$-additivity)

then, $(\Omega, \mathcal{F}, P)$ is a probability space.

Examples of event space

Fair die: $\Omega_1 = \{\omega_1, \ldots, \omega_6\}$ with $P(\omega_i) = \frac{1}{6}$,

Fair coin: $\Omega_2 = \{\psi_1, \psi_2\}$ with $P(\psi_j) = \frac{1}{2}$,

Product spaces:

$\Omega = \Omega_1 \times \Omega_2$, or

$\Omega = \Omega_1 \times \Omega_1$

Bernoulli trials: $\underbrace{\Omega_2 \times \Omega_2 \times \cdots \times \Omega_2}_{n \text{ times}}$.

ω_i = sum of 2 independent dice,

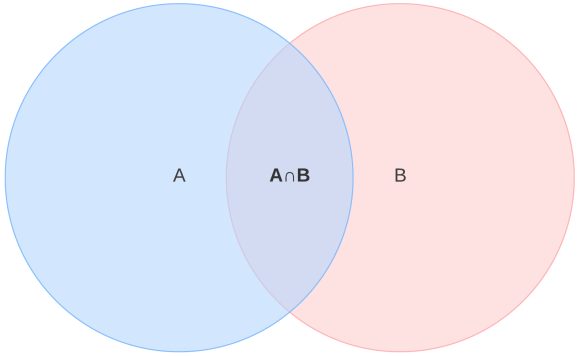
$\Omega = \{2, 3, \ldots, 12\}$.

Basic properties

0. Unity:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Venn Diagram



1. Independence*:

$$A, B \text{ independent events} \rightarrow P(A \cap B) = P(A) \cdot P(B).$$

2. Conditional Probability:

$$P(B) \neq 0 \rightarrow P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

$$(A, B \text{ independent} \rightarrow P(A|B) = P(A)).$$

*Independence: "The occurrence of one event does not affect the probability of occurrence of the other."

Basic properties

3. Chain Rule:

$$P(A_1 \cap \cdots \cap A_n) = P(A_n | A_{n-1}, \ldots, A_1) \cdot P(A_2 | A_1) \cdot P(A_1)$$
$$= \prod_{k=1}^{n} P(A_k | \cap_{j=1}^{k-1} A_j).$$

4. Bayes' Theorem:

If $P(B) \neq 0$ then $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$.

What is a Random Variable?

- A *Random Variable* (r.v. or rv) is a function from Ω to $\mathbb{R}$:

$$\omega \to x(\omega)$$

e.g., ω_i: face of a die, $x(\omega_i) = 10i$, voltage of a random source, etc.

- We'll use the notation: $x(\omega), X(\omega), Y(\omega), ...$ or just $x, X, Y, ...$

Basic properties of r.v.s

- Cumulative Density Function (CDF), or Distribution Function:

$$F_X(x) := P\,(X \leq x) \equiv P\,(\{\omega \in \Omega : x\,(\omega) \leq x\}).$$

- $\lim_{x \to -\infty} F_X(x) = P(\emptyset) = 0.$
- $\lim_{x \to +\infty} F_X(x) = P(\Omega) = 1.$

Basic properties of r.v.s

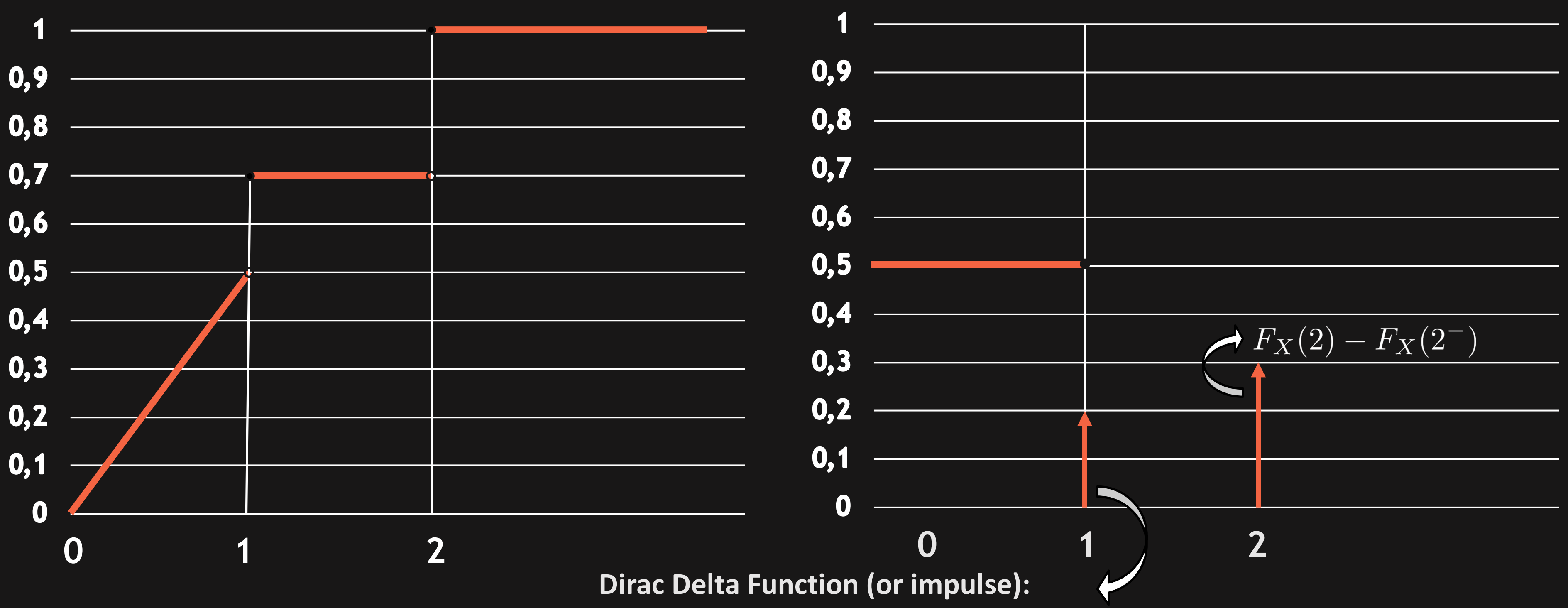
- Probability Density Function (PDF):

$$f_X(x) = \frac{d}{dx} F_X(x).$$

- $f_X(x) \geq 0$, because $F_X(x)$ is an increasing function of x.
- $\int_{-\infty}^{+\infty} f_X(x) dx = 1.$

Example of a CDF/PDF

- <u>Mixed-type PDF:</u>



Dirac Delta Function (or impulse):

$$\delta(x) : \int_{-\infty}^{+\infty} \phi(x)\delta(x)dx = \phi(0).$$

Moments of a r.v.

- <u>Expected Value:</u>

$$\mu_X = \mathbb{E}_X[X] = \int_{-\infty}^{+\infty} x f_X(x) dx$$

"center of mass"

$$\approx \frac{1}{n}\sum_{i=1}^{n} x_i \,,\ x_i \sim f_X,\ \text{i.i.d.}^*$$

empirical

*i.i.d: "Independent and identically distributed random variables."

Moments of a r.v.

- <u>Variance:</u>

$$\sigma_X^2 = \mathbb{E}_X\left[(X-\mu_X)^2\right]$$
$$= \int_{-\infty}^{+\infty} (x-\mu_X)^2 f_X(x)dx$$
$$= \mathbb{E}_X\left[X^2\right] - \mu_X^2.$$

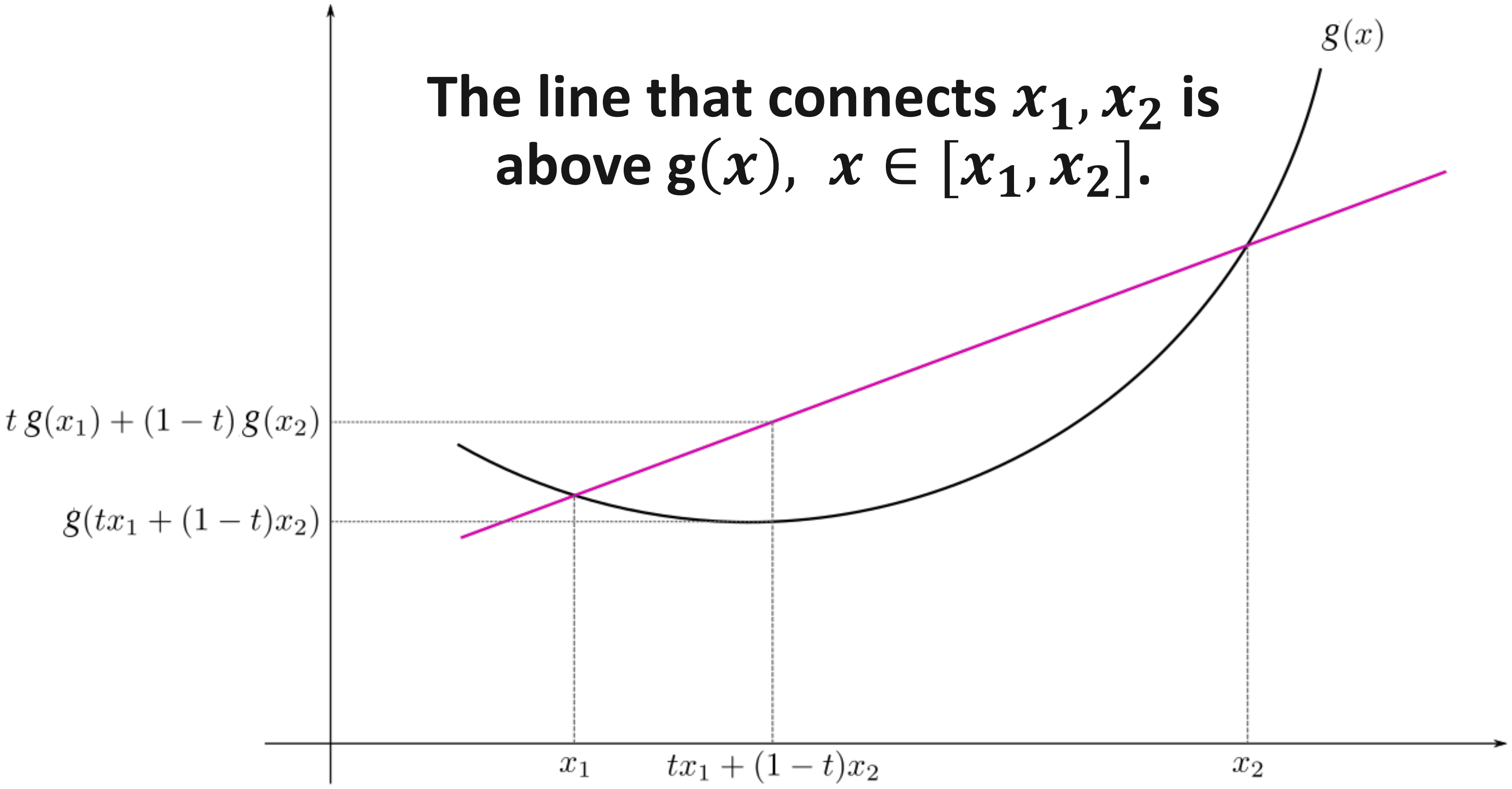
$$\mathbb{E}_X[g(X)] = \int_{-\infty}^{+\infty} g(x) f_X(x)dx \neq g(\mathbb{E}_X[X]).$$

- <u>Standard Deviation:</u>

$$\sigma_X = \sqrt{\sigma_X^2}.$$

Jensen's Inequality

- Convexity: $g(\cdot)$ convex $\iff g(tx_1 + (1-t)x_2) \le tg(x_1) + (1-t)g(x_2)$



- <u>Jensen's Inequality:</u>

$$g(\cdot) \text{ convex}$$

$$\Rightarrow g\left(\mathbb{E}[X]\right) \le \mathbb{E}\left[g(X)\right]$$

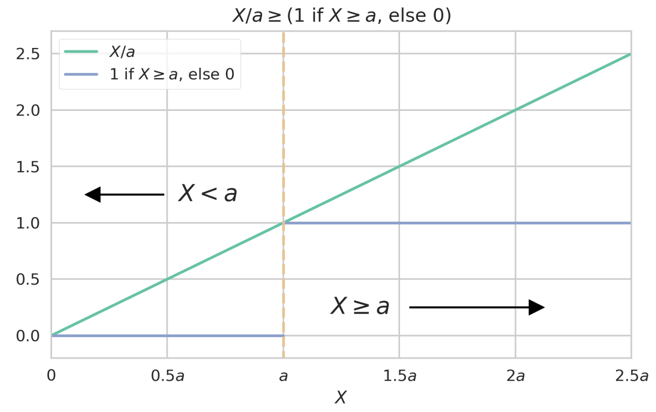
Markov's Inequality

- Let X be a non-negative r.v., and $a > 0$. Then,

$$P(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

- ***Proof:***

$$\begin{aligned}\mathbb{E}[X] &= \int_{-\infty}^{+\infty} x f(x) dx \\ &= \int_{0}^{+\infty} x f(x) dx \\ &= \int_{0}^{a} x f(x) dx + \int_{a}^{+\infty} x f(x) dx \\ &\geq \int_{0}^{a} x f(x) dx + \int_{a}^{+\infty} a f(x) dx \\ &\geq a \int_{a}^{\infty} f(x) dx \\ &= a P(X \geq a).\end{aligned} \Bigg\} \implies P(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$



Chebyshev's Inequality

- Let X be a r.v. with finite expected value μ, and finite non-zero variance σ^2. Then, for any positive real number $k > 0$:

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

Proof: Apply Markov's inequality to $Y = (X - \mu)^2$.

- E.g., for $k = 2$:

$$P(|X - \mu| \leq 2\sigma) \leq \frac{1}{4} = \frac{25}{100}, \ \forall \text{r.v. } X \text{ with } \mu, \sigma < \infty.$$

  - When X is Gaussian, it holds: $P(|X - \mu| \leq 2\sigma) \leq \frac{5}{100}$.

Sampling an r.v. given $F_X(x)$

- Probability Integral Transform:

If X is a continuous r.v. with CDF $F_X(x)$, then the r.v. $Y = F_X(X)$ has a uniform distribution in $[0,1]$:

$$\left.\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(F_X(X) \leq y) \\ &= P(X \leq F_X^{-1}(y)) \\ &= F_X(F_X^{-1}(y)) \\ &= y,\ y \in [0,1]. \end{aligned}\right\} \implies f_Y(y) = \begin{cases} 1, & 0 \leq y \leq 1 \\ 0, & \text{elsewhere.} \end{cases}$$

assuming $F_X^{-1}(y)$ exists

Sampling an R.V. given $F_X(x)$

- Inverse version (Inverse Transform Sampling):

If Y has a uniform distribution in $[0,1]$ and X has CDF $F_X(x)$, then the r.v. $F_X^{-1}(Y)$ has the same distribution as X.

Very important and popular example: categorical distribution or softmax.

- Algorithm:

1. Compute the inverse of $F_X(x)$, i.e., $F_X^{-1}(x)$.
2. Generate a random number $u \sim U([0,1])$.
3. Compute $x = F_X^{-1}(u) \sim X$.

Function of an r.v.

- Let $g(\cdot)$ be a monotonic function. The r.v. $Y = g(X)$ has PDF given by:

$$f_Y(y) = f_X(g^{-1}(y)) \cdot \left|\frac{d}{dy} g^{-1}(y)\right| = f_X(g^{-1}(y)) \cdot \left|\frac{1}{g'(g^{-1}(y))}\right|.$$

Proof:

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(g(X) \leq y) \\ &= P(X \leq g^{-1}(y)) \\ &= F_X(g^{-1}(y)). \end{aligned}$$

Examples:

$Y = aX + b,\ a > 0,$

$f_Y(y) = \frac{1}{a} f_X\left(\frac{y-b}{a}\right),$

$Y = X^3,$

$f_Y(y) = \ldots$

Joint r.v.s

- Joint CDF of 2 random variables:

$$P(X \leq x, Y \leq y) = F_{XY}(x, y).$$

Limits:

$$P(x, +\infty) = F_X(x),$$

$$P(+\infty, y) = F_Y(y).$$

PDF:

$$f_{XY}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{XY}(x, y) \geq 0.$$

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f_{XY}(x, y) dx dy = 1.$$

Independence

- Independence of X, Y:

$$P(X \leq x, Y \leq y) = P(X \leq x) \cdot P(Y \leq y)$$
$$\Longrightarrow F_{XY}(x, y) = F_X(x) \cdot F_Y(y)$$
$$\Longrightarrow f_{XY}(x, y) = f_X(x) \cdot f_Y(y).$$

Sum of two *independent* r.v.s

- Let X, Y be two independent random variables. The PDF of $Z = X + Y$ is given by the **convolution** of $f_x(z)$ with $f_y(z)$:

$$\left.\begin{aligned} P(Z \leq z) &= P(X + Y \leq z) \\ &= P(X \leq z - Y) \\ &= \int_{-\infty}^{+\infty} P(X \leq z - y) f_Y(y) dy \\ &= \int_{-\infty}^{+\infty} F_X(z - y) f_Y(y) dy \end{aligned}\right\} \Longrightarrow$$

$$\begin{aligned} \Longrightarrow f_Z(z) &= \int_{-\infty}^{+\infty} f_X(z - y) f_Y(y) dy \\ &= f_X(z) * f_Y(z). \end{aligned}$$

Examples: Sum of 2 uniform distributions, sum of 2 dice, etc.

Conditional PDFs

- **Conditional Probability:** *Example:* Die with even/odd events as conditions

$$f_{Y|X}(y|X=x) = f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)} = \frac{f_{XY}(x,y)}{\int_{-\infty}^{+\infty} f_{XY}(x,y)dy}$$

- **Marginal Probability:**

$$f_Y(y) = \int_{-\infty}^{+\infty} f_{Y|X}(y|x) f_X(x) dx$$

- **Bayes Theorem:**

$$f_{Y|X}(y|x) = \frac{f_{X|Y}(x|y) f_X(x)}{f_Y(y)}$$

Both a consequence of:

$$f_{XY}(x,y)$$

$$= f_{Y|X}(y|x) f_X(x)$$

$$= f_{X|Y}(x|y) f_Y(y)$$

Multivariate Gaussian Distribution

- **Notation:**

$$x = (x_1, \ldots, x_d)^T, \quad x \sim \mathcal{N}(\mu, \Sigma).$$

- **PDF General Form:**

$$\begin{aligned} \mathcal{N}(x|\mu, \Sigma) &= f_X(x_1, \ldots, x_d) \\ &= \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}, \quad |\Sigma| \triangleq \det(\Sigma). \end{aligned}$$

$$|\Sigma| \neq 0 \rightarrow \mathcal{N}(x|\mu, \Sigma) \text{ non-degenerate.}$$

Multivariate Gaussian Distribution

- Mean Vector:

$$\mu = \mathbb{E}[X] = (\mathbb{E}[X_1], \mathbb{E}[X_2], \ldots, \mathbb{E}[X_d])^T$$

$$= \int_{-\infty}^{+\infty} x f_X(x|\mu, \Sigma) dx \in \mathbb{R}^d.$$

- Covariance Matrix:

$$\Sigma_{ij} = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)] = \mathrm{Cov}[X_i, X_j] \in \mathbb{R}^{d \times d}.$$

Multivariate Gaussian Distribution

- Marginals (2D example):

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}.$$

$$p(x_1) = f_{X_1}(x_1) = \int f_X(X|\mu, \Sigma) dx_2 = \mathcal{N}(x_1|\mu_1, \sigma_1^2).$$

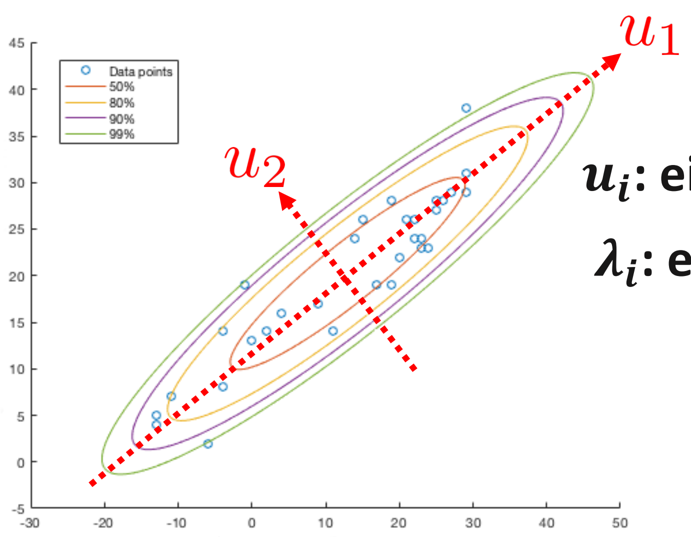
$$p(x_2) = ... = \mathcal{N}(x_2|\mu_2, \sigma_2^2).$$

Multivariate Gaussian Distribution

- <u>Geometric interpretation:</u>

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}, \quad \rho = \frac{\sigma_{12}}{\sigma_1\sigma_2}.$$

Correlation coefficient



u_i: eigenvectors

λ_i: eigenvalues

$$\left.\begin{aligned} \Delta^2 &= (x-\mu)^T\Sigma^{-1}(x-\mu) \\ \Sigma^{-1} &= \sum_{i=1}^{d} \frac{1}{\lambda_i} u_i u_i^T \end{aligned}\right\} \Longrightarrow$$

$$\Delta^2 = \sum_{i=1}^{d} \frac{y_i^2}{\lambda_i}, \quad y_i = u_i^T(x-\mu).$$

Multivariate Gaussian Distribution

- <u>How to sample:</u>

$$x = \mu + \sigma z \sim \mathcal{N}(\mu, \sigma^2)\ , \quad z \sim \mathcal{N}(0,1)$$

Cholesky decomposition

Reparametrization trick

$$x = \mu + Lz \sim \mathcal{N}(\mu, \Sigma)\ , \quad z \sim \mathcal{N}(0, I_d) \quad \text{and } \Sigma = LL^T.$$

<u>Sampling from 1d normal</u>

```
import numpy as np
mu, sigma = 0, 0.1 # mean and standard deviation
x = np.random.normal(mu, sigma, 1000)
x.shape → (1000,)
type(x) → <class 'numpy.ndarray'>
```

<u>Sampling from multivariate normal</u>

```
mu = [1, 2]
Sigma = [[1, 2], [2, 4]]
x = np.random.multivariate_normal(mu, Sigma, 1000)
x.shape → (1000,2)
```

Multivariate Gaussian Distribution

- <u>Conditional probability:</u>

$$x = \begin{bmatrix} x_A \\ x_B \end{bmatrix} \qquad x_A = \begin{bmatrix} x_1 \\ \vdots \\ x_{d_1} \end{bmatrix} \qquad x_B = \begin{bmatrix} x_{d_1+1} \\ \vdots \\ x_d \end{bmatrix}$$

$$\mu = \begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}.$$

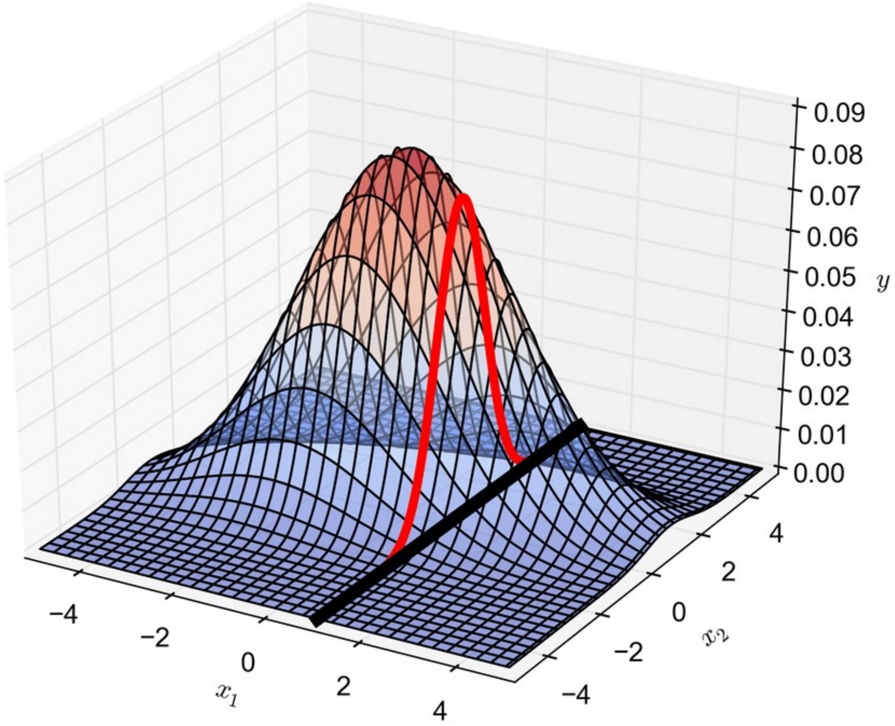
$$p(x_A|x_B) = \mathcal{N}(x_A|\mu_A + \Sigma_{AB}\Sigma_{BB}^{-1}(x_B - \mu_B), \underbrace{\Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}}_{\text{Schur complement}})$$

$$p(x_B|x_A) = \mathcal{N}(x_B|\mu_B + \Sigma_{BA}\Sigma_{AA}^{-1}(x_A - \mu_A), \Sigma_{BB} - \Sigma_{BA}\Sigma_{AA}^{-1}\Sigma_{AB})$$

Multivariate Gaussian Distribution

- Conditional probability example:

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}$$



$$p(x_2|x_1) = \mathcal{N}(x_2|\mu_2 + \sigma_{21}\sigma_1^{-2}(x_1 - \mu_1),\ \sigma_2^2 - \sigma_{21}^2\sigma_1^{-2})$$

– Variance (a.k.a. uncertainty) is reduced whenever there is correlation!

Asymptotics

- *Law of Large Numbers:*

Let $\{X_k\}$ be a sequence of i.i.d. r.v.s and $S_n = \sum_{k=1}^{n} X_k$ be the sum r.v.
If $\mu = \mathbb{E}[X_k]$ exists, then $\forall \epsilon > 0$:

$$\lim_{n\to\infty} P\left(\left|\frac{1}{n}S_n - \mu\right| > \epsilon\right) = 0.$$

E.g., a fair coin, fair dice, etc.

Proof: Application of Chebyshev's inequality.

Asymptotics

- *Central Limit Theorem:*

Let $\{X_k\}$ be a sequence of i.i.d. r.v.s and $S_n = \sum_{k=1}^{n} X_k$ be the sum r.v. Suppose that $\mu = \mathbb{E}[X]$, and $\sigma^2 = \mathrm{var}(X)$ exist. Then, $\forall \beta$ fixed:

$$\lim_{n\to\infty} P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} < \beta\right) = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{\beta}{\sqrt{2}}\right)\right),$$

where $\mathrm{erf}(\cdot)$ is the error function defined by: $\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt.$

- In other words: $\frac{S_n - n\mu}{\sigma\sqrt{n}} \sim \mathcal{N}(0,1)$, as $n \to \infty$.

References

1. All of statistics: A Concise Course in Statistical Inference (*Chapters 1–4*)
   Larry Wasserman, Springer (2004)

2. Probabilistic Machine Learning: An Introduction (*Chapters 2–3*)
   Kevin P Murphy, The MIT Press (2022)

Introduction to Deep Generative Modeling

Lecture #2

HY-673 – Computer Science Dep., University of Crete

Professors: Yannis Pantazis & Yannis Stylianou

TAs: Michail Raptakis & Michail Spanakis